Introduction to Next Generation Sequencing Technology for Non Practitioners

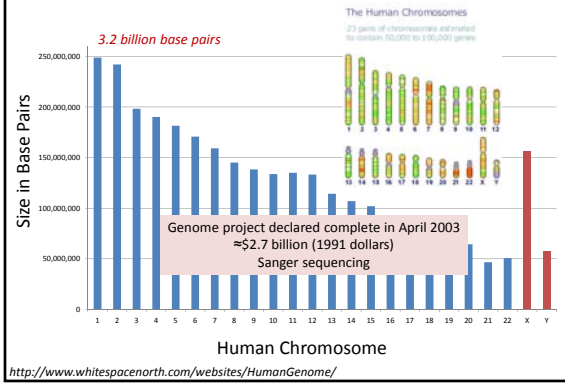
Peter M. Vallone PhD
Leader, Applied Genetics Group
NIST
National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Global Identity Summit
DNA: Next Generation Sequencing Technology Application to Human ID
September 22, 2015

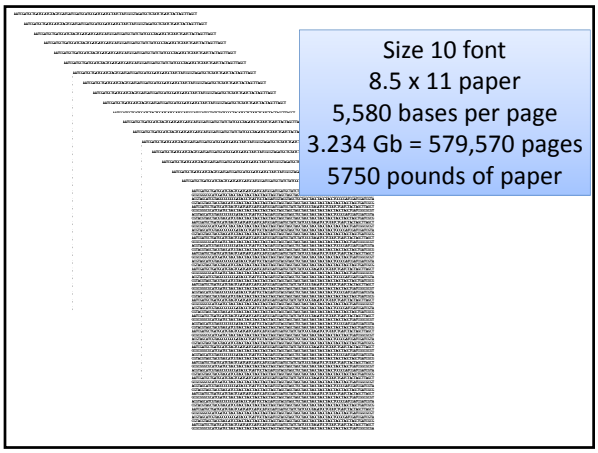
Outline

- Human Genome
- Sanger sequencing
- 'Disruptive innovation'
- Non-identity applications
- Generalized workflows for NGS
- Platforms and throughput
- Sequencing chemistries
- Wrap up / thoughts

Human Genome



http://www.whitespacenorth.com/websites/HumanGenome/



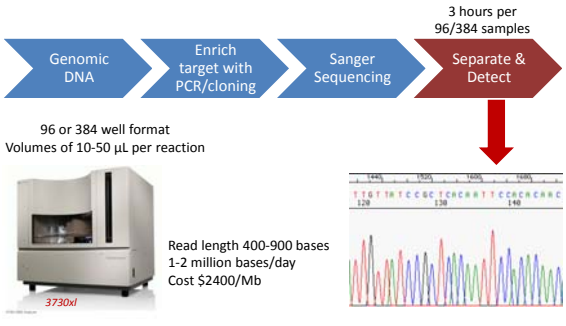
Sanger Sequencing

How the human genome was first completed April 2003

- 3.2 billion bp
  - Each chromosome 50 to 300 million bp
  - Clone 150-200 million bp pieces into bacteria (BAC)
  - Store fragments and replicate
  - 20,000 BACs for the human genome
  - Each BAC is cut into 2,000 bp
  - Sequenced 500-800 bp (10x) depth
  - Map to chromosome and assemble…

Generalized Sanger Workflow

Fluorescently labeled DNA fragments are separated and detected by capillary electrophoresis



What's in a name???

Massively parallel sequencing

NGS

Second-generation sequencing

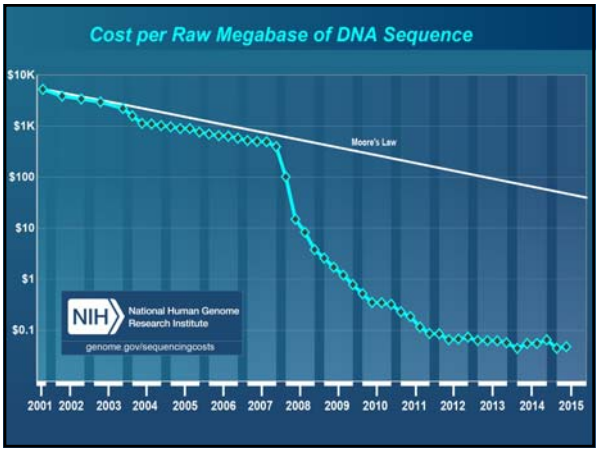
Next-generation sequencing

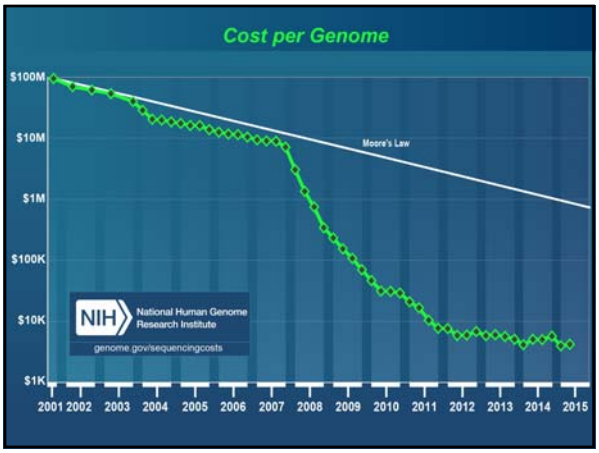
Whole-genome sequencing

Third-generation sequencing

HIGH-THROUGHPUT SEQUENCING

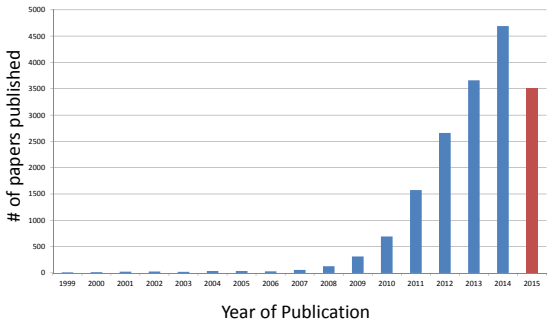
Next-generation genomics





PubMed Search results for "Next Generation Sequencing"

http://www.ncbi.nlm.nih.gov/pubmed



Non-identity applications

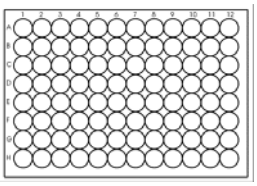
- Clinical
- Inherited disease
- Reproductive health
- Cancer – gene fusion
- Rare variants
- Pre-implantation (genetic screening)
- Transplant medicine (HLA)
- Microbiomics/Metagenomics
- Gene expression | RNA seq
- Public health
- Ancient DNA
- NIPT (non-invasive prenatal testing)

Next Generation Sequencing

Massively Parallel Sequencing

'Millions of wells'

- One DNA fragment per 'well'
- Typically, shorter reads (Range 75 to 400)
- High coverage 100 – 1000 - 10,000x
- **Rely more on informatics to assemble millions of short reads**

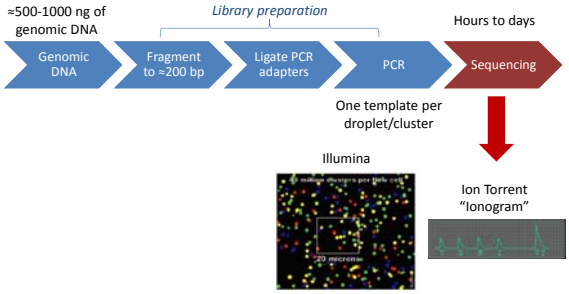


- Each well of the plate will generate 500-900 bp of sequence data
- One plate = 86,400 bp of sequence data
- 3 h per run (max 8 runs/day)
- 691,200 (≈ 0.7 mb per day)

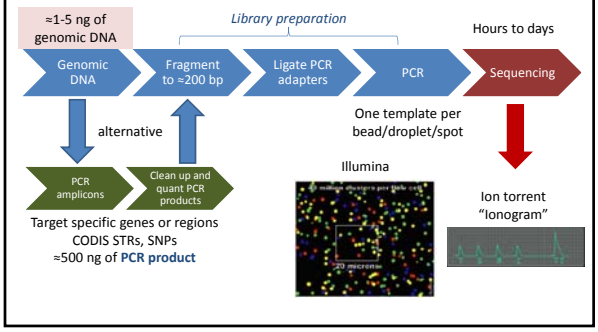
- **Millions of 'wells' generating 200-400 bp of sequence data**
- **nL of reagents per reaction**
- **2 hours to 6 days per run**
- **0.5 - 1800 Gb of sequence data per run**

Massively parallel sequencing takes advantage of partitioning of molecules

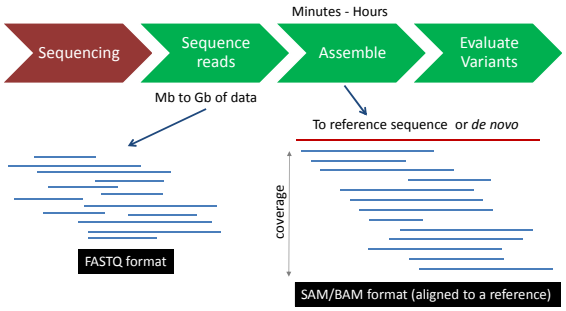
Generalized NGS Workflow



Generalized NGS Workflow



Generalized NGS Workflow



FASTQ Format

- FASTQ - normally uses four lines per sequence.
- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.
- (1)@SEQ_ID
  (2)GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCC
  (3)+
  (4)!''*((((***+))%%%++)(%%%%).1***-+*''))**

http://maq.sourceforge.net/fastq.shtml

Informatics

After the generation of sequence data

- Storage of data
- Assembly and parsing of data
- Variant calling
- Applied use of information
- Hire a bioinformatics expert…

Sampling of NGS Platforms

- Illumina
  - **MiSeq**
  - **FGx**
  - HiSeq 2000/2500
  - NextSeq
- Life Technologies
  - SOLiD (5500 series)
  - **Ion Torrent PGM**
  - Ion Torrent Proton
- Pacific Biosciences
  - PACBIO RS II
- Oxford Nanopore
  - MinION
- 454 Roche
  - GS jr
  - GS FLX+





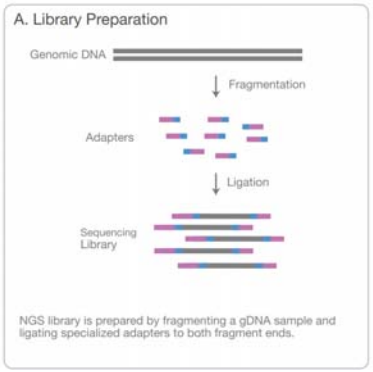
October 15, 2013 – Roche shutting down 454 sequencing business
Will be phased out mid-2016

NGS Chemistry

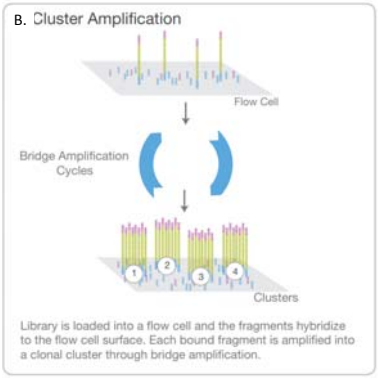
- Not Sanger sequencing
- Examples
  - Sequencing by synthesis (Illumina)
  - Ion semiconductor sequencing (Ion Torrent)
  - Single molecule real time (Pacific Bioscience)
  - Nanopore sequencing (Oxford Nanopore)

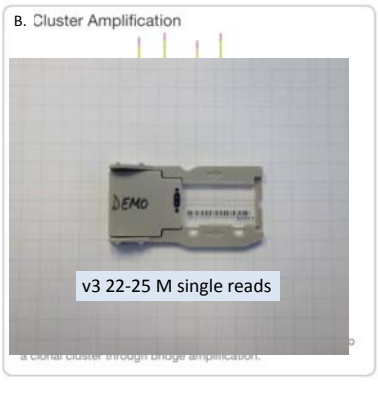
Illumina MiSeq

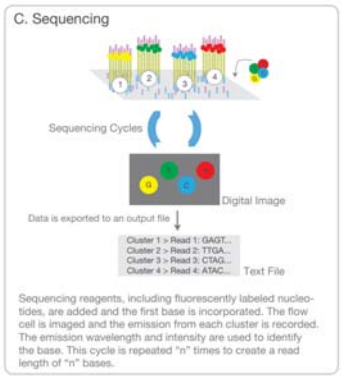
- MiSeq launched in Jan. 2011
- The MiSeq uses a sequencing by synthesis approach:
  - Nextera enzymatically fragments and tags DNA
  - Limited cycle PCR
  - Flow cell hybridization
  - Bridge PCR - clusters
- Fluorescent light detection
  - Each base has a unique color
  - Sequence each end of the molecule



http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf



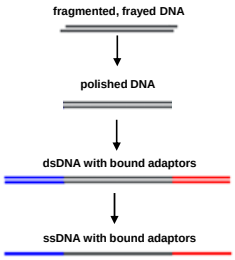




Life Tech - Ion Torrent - PGM

- Ion Torrent launched in Feb. 2010
- Ion Torrent sequencing employs an analogous technique as pyrosequencing:
  - Emulsion PCR for single copy reactors
  - Non-labeled nucleotide triphosphates are flowed over a bead on a semiconductor surface
- Hydrogen Ion detection
  - pH change is detected
  - **No optics**

Constructing a Library

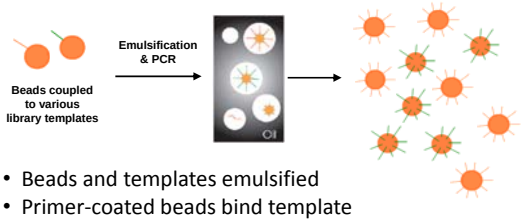


- Frayed ends are enzymatically "polished"
- Adaptor oligos ligated onto fragments
- Denatured to ssDNA
- Emulsified with primer-coated beads
- Hybridization of template to bead

Margulies et al. (2005) Supplementary Materials

Ion Torrent - PGM

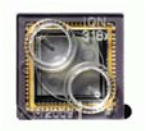
Emulsion PCR & Enrichment



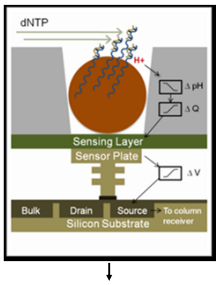
- Beads and templates emulsified
- Primer-coated beads bind template
- PCR amplifies template
- Enrich for beads containing PCR products
  - magnetic capture
- Adaptable to automation (Ion Chef)

Dressman et al. (2003)

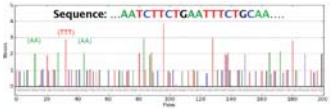
Ion torrent PGM chip



'318' chip
11 M sensors



- Chip flooded with one nucleotide after another
- H^+ released when a complementary base is added to template
- Charge from the ion causes detectable pH change
- Sequencer calls the base



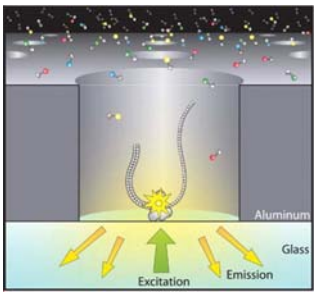
Life Technologies

Ion Torrent - PGM

Pacific Biosciences

- Instrument = PacBio RS II (2011)
- SMRT = single molecule real time
- NTPs are attached to an unique fluorescent dye
  - **A**, **C**, **G**, **T**
- Sequence is 'read' as the template is being synthesized (in real time)
- Long reads >14,000 bp

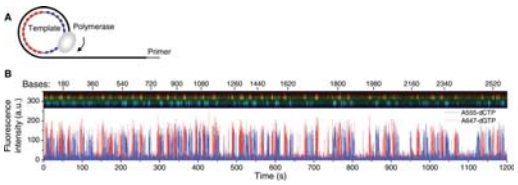
SMRT = single molecule real time



- DNA polymerase is immobilized at the bottom of the well
- Fl-labeled ACGT diffuse into the well and are incorporated
- Light is detected as dNTPs are incorporated

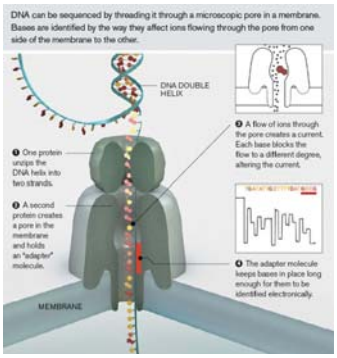
http://chrisamiller.com/science/2010/02/10/third-gen-sequencing-pacific-biosciences/
Eid, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science doi:10.1126/science.1162986

SMRT = single molecule real time



http://www.sciencemag.org/content/323/5910/133/F3.large.jpg

Nanopore Sequencing



http://www2.technologyreview.com/article/427677/nanopore-sequencing/

Nanopore Sequencing

http://www2.technologyreview.com/article/427677/nanopore-sequencing/

NGS Specs as of Sept 2015

	NGS Platform	Capacity/Output	Read Length	Approx run time
ThermoFisher	Ion PGM	30 Mb to 2 Gb (1.2-11 M)	200-400 bp	2 hours
	Ion Proton	Up to 10 Gb (165-660 M)	200 bp	2-4 hours
Illumina	MiSeq Series	15 Gb (25 M)	2 x 300 bp	5-65 h
	HiSeq Series	1500 Gb (5000 M)	2 x 150 bp	40 h to 6 days
	NextSeq Series	120 Gb (400 M)	2 x 150 bp	12-30 h
	HiSeq X Series	1800 Gb (6000 M)	2 x 150 bp	3 days
Pacific Biosciences	PACBIO RS II	Up to 1 Gb (150 k)	> 20 kb	0.5 to 4 h

[1] https://www.illumina.com/content/dam/illumina-marketing/documents/products/brochures/brochure_sequencing_systems_portfolio.pdf
[2] http://files.pacb.com/pdf/PacBio_RS_II_Brochure.pdf
[3] https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-proton-system-for-next-generation-sequencing/ion-proton-system-specifications.html

Moving Targets

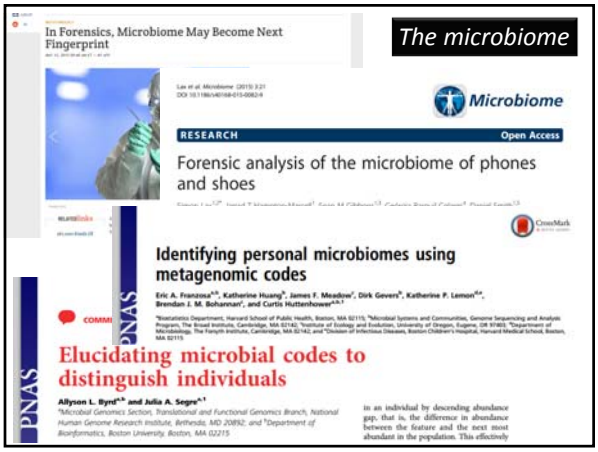
6 months from now these parameters will have changed

- Newer instruments
- Costs decreasing
- Throughput increasing
- Read lengths increasing
- Chemistries improving
- Library preparations – simpler/automated
- Computers faster – data storage cheaper
- Platforms leaving the market (e.g. Roche 454)
- Platforms entering the market (e.g. Qiagen GeneReader, Oxford Nanopore MinION)

Use of NGS for identity applications

Highly-parallel/high-throughput direct sequencing of relevant targets

- Forensic and/or biometric genetic markers
  - newer human identity applications
  - biogeographical ancestry, externally visible traits, complex kinship
  - degraded samples, mixtures, low template
- Traditional Forensic Markers
  - Whole mitochondrial genome analysis
  - Going in depth into STR loci and beyond



General Comments

- NGS is a 'disruptive innovation'
- NGS platforms/technologies open up high throughput access to genomic information
  - Whole genome or targeted regions
- The amount of sequence data generated makes the technique economical
  - Information per unit time
  - Current cost: days and thousand of dollars per run
  - BUT always decreasing and automation is helping
- Bioethics?

Not rapid/'field ready'

Thanks for your attention!

Questions and discussion?

Peter.Vallone@nist.gov
301-975-4872

NIST Disclaimer: Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

Outside funding agencies:
FBI - Evaluation of Forensic DNA Typing as a Biometric Tool